DOCUMENT RESUME

ED 270 480 TM 860 362

AUTHOR Hwang, Chi-en; Cleary, T. Anne
TITLE Comparing the IRT Pre-equating and Section Pre-equating: A Simulation Study.
PUB DATE Apr 86
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Simulation; *Equated Scores; *Latent Trait Theory; Mathematical Models; *Sample Size; Statistical Studies; *Test Construction; Test Items; *Test Length
IDENTIFIERS *Pre Equating (Tests); *Section Pre Equating (Tests)

ABSTRACT

The results obtained from two basic types of pre-equatings of tests were compared: the item response theory (IRT) pre-equating and section pre-equating (SPE). The simulated data were generated from a modified three-parameter logistic model with a constant guessing parameter. Responses of two replication samples of 3000 examinees on two 72-item tests were generated. Tests of various numbers of non-operational items were simulated by ignoring certain item responses of the examinees. Smaller samples were drawn from the large pool. The analyses involved equating the two tests with five different lengths of non-operational test sections (12 to 48 items) and six different sample sizes (300 to 3000) using both the IRT and SPE methods. Results suggested greater accuracy for both IRT and SPE when sample sizes were larger than 1500 and longer tests were used. For smaller samples and shorter tests, however, the IRT results were less satisfactory than those of SPE. The results suggested that SPE may provide more accurate equating than the IRT method when small samples (under 1000) and shorter tests (under 90 items) are necessary. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED270480

Comparing the IRT Pre-equating and Section Pre-equating: A Simulation Study

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Chi-en Hwang

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Chi-en Hwang

and

T. Anne Cleary

The University of Iowa

For presentation at the Annual Meeting of the
American Educational Research Association,
San Francisco, April 1986

TM 860 362

Abstract

This study compared the results obtained from two basic types of pre-equatings: the item response theory (IRT) pre-equating and section pre-equating (SPE). The data used were generated from a modified three-parameter logistic model with a constant guessing parameter. Responses of two replication samples of 3000 examinees on two 72-item tests were generated. Tests of various numbers of non-operational items were simulated by ignoring certain item responses of the examinees. Smaller samples were drawn from the large pool. The analyses involved equating the two tests with five different lengths of non-operational test sections (12, 18, 24, 36, and 48 items) and six different sample sizes (300, 600, 900, 1500, 2250, and 3000 examinees) using both the IRT and SPE methods. Pre-equating results were evaluated against an IRT criterion equating using complete data, and the standardized root mean square differences between pre-equatings and the criterion were used to evaluate the equatings. There was a slight trend toward greater accuracy for both IRT and SPE when sample sizes were larger than 1500 and longer tests were used. For smaller samples and shorter tests; however, the IRT method failed to converge, and the obtained results were less satisfactory than those of SPE. The results of this study suggested that SPE may provide more accurate equating than the IRT method when small samples and shorter tests are necessary.

COMPARING THE IRT PRE-EQUATING AND SECTION PRE-EQUATING: A SIMULATION STUDY

Introduction

The purpose of this study was two-fold: First, to compare the equating results obtained by item response theory pre-equating methods with those obtained by section pre-equating (SPE) methods. Second, to compare the effects of sample size and test length on pre-equating results both within and across the two methods.

Many testing programs have several administrations every year, and multiple forms of a test are required so that different forms can be used in each administration. Multiple forms are widely used in many test situations such as admission tests, achievement tests and licensing and certification tests. The use of alternate forms assures that information about specific items on the test cannot be made available to test takers at a later administration, giving those test takers an unfair advantage over the people who took the same test on a earlier administration. Thus, alternate forms are needed to maintain test security.

Accompanying the use of multiple or alternate forms is the issue of equating. Since two forms of a test can rarely be made precisely equal in difficulty, it becomes necessary to convert from the units of one form of a test to the units of the other so that scores derived after conversion will be equivalent. This process of conversion,

usually by means of some kind of statistical adjustment of test scores, is commonly called test equating. The product of conversion may involve either an equation or a table of converted scores associated with each observable score.

Conventionally, equatings have been done by including a set of items called anchor items from a previously administered form (Angoff, 1984). Because the anchor items are taken by both the current group and the previous group of examinees, they provide a link between the current form and the previous form. This link makes it possible to convert scores on the new form to the scale from the previous form. The conventional equating design and practice, however, is challenged by the test disclosure legislation. Recent legal requirements for public disclosure of test questions once they have been used operationally have made it difficult to maintain test security. There is also voluntary disclosure in many national testing programs where disclosure legislation does not exist. In response to the problems presented by test disclosure, pre-equating methods have been developed. In pre-equating, a new test form is equated to an old form prior to the actual administration of the new form. Pre-equating methods permit the use of different spiralled equating sections to different subsamples of examinees instead of a fairly large set of items to all examinees. The final form of the new test is drawn from these precalibrated samples and is equated before its actual administration. Because the pre-equating items are not used for reporting scores, they need not to be disclosed. The security of the new form will thus be maintained.

Another advantage of pre-equating is the convenience of obtaining the needed equating equations before the operational administration of a form, thus shortening the time required to prepare score reports after a test administration (Cowell, 1982; Holland & Wightman, 1982; Cook & Eignor, 1983). When the new test form consists entirely of precalibrated items, the equating can be done before the test is given, and the period between test administration and score reporting can be shortened.

There are basically two widely used methods of pre-equating: one uses the methods of item response theory (Lord, 1980) and the other is known as section pre-equating or SPE (Holland & Wightman, 1982). Item response theory assumes that the relations between the probability of a correct response on an item and the ability of an examinee can be represented by a mathematical function. In this model, the probability of a correct response to a given item by an examinee of a specified ability is a function of a vector of parameters describing the item. In the three-parameter IRT model, the parameter b_i represents the difficulty of the item i; a_i is the slope parameter at the point of inflection and represents the item discrimination; and c_i is the lower asymptote of the function and represents a guessing parameter.

Most IRT equating is currently attempted by the true-score equating procedure described by Lord (1980). If the ability level of an examinee and the item parameters for all the items in a test were known, the expected true score of an examinee equals the sum of the probabilities of correctly answering each item. Suppose two tests are

measuring the same ability, the corresponding true scores have identical meaning and they are called equated true scores. In practice, estimated values of item parameters described above are used to calculate the expected true scores for tests being equated. Item response theory has the advantage for score equating that the item parameters have known relationships to the test scores. If items from several sources are pretested in a common population, item parameter estimates can be developed to form a large pool of precalibrated items. New forms can be built from this pool and can be equated before they are administered.

Section pre-equating methods derive from conventional linear equating, but utilize somewhat more detailed designs for data collection and algorithms for estimating parameters of the linear transformations. Section pre-equating administrations typically consist of separately timed sections with most sections designated for scoring referred as operational sections. Other sections are used for pre-equating and are called variable sections. Different sections of the new form are given to random samples of examinees who take the entire old form. The means, variances and covariances of the scores of the new form are estimated from samples given different variable sections. A linear function is then applied to these estimates to equate the new form to the old form.

The basic problem for SPE is to estimate the correlation, $\rho(y_i y_j)$, between variable sections when only partial information is available. Given that a new form, Y, is to be equated to an old form, X, through SPE, the estimation of means and variances for the two

tests are required for the linear equating function. Holland and Wightman (1982) have discussed several methods and situations to which they can be applied. The simplest way to estimate the mean and covariance matrix of Y is by pairwise presentation, wherein each sample from P takes two non-operational(or variable) sections from the new form Y. We can use the data from the random sample P_{ij} that is tested with section Y_i and Y_j to calculate the estimate of population covariance, $\sigma_{y_i y_j}$, and pool all the samples tested with Y_i to compute the unbiased estimate of population mean and variance of Y_i. The population estimate of the correlation, $\rho(y_i y_j)$, can be obtained by combining the estimates of the covariances and variances . The restriction of pairwise presentation is that it only applies to the two-variable-section test (2VST) situation.

In the one-variable-section test (1VST) where groups of examinees share no common items in variable sections, the estimation process of covariances between variable sections becomes more complicated. Several methods of estimating the missing correlation between two variable sections were discussed in Holland and Wightman (1982) including borrowing correlations from another population in which Y is operationally administered, borrowing the correlations among X-sections as estimates of correlations among Y-sections, and estimating parameters using the EM algorithm. The EM method (Holland & Thayer, 1984) assumes a multivariate normal distribution over the population and uses maximum-likelihood estimation for the covariance matrix that would have been observed had examinees taken all sections of either Form X and Form Y. The EM algorithm estimation, being more compli-

cated and more expensive than the other methods, has been claimed to produce estimates closer to the true population value than the pairwise method because it makes use of all of the observable data.

Results from both SPE and IRT pre-equating methods have been compared with results from a variety of equating methods including linear, equipercentile, and IRT equatings. However, these two methods used for pre-equating test scores have not been directly tested under exactly parallel circumstances until recently. The Leary and Wightman (1985) study was the first empirical study directly comparing results of IRT equating with those obtained by section pre-equating using pairwise presentation method and those obtained by section pre-equating using the EM algorithm for parameter estimations. In that study, two parallel new forms were pre-equated to each other by the three methods mentioned above. The three models were evaluated against a criterion based on the consistency of equating between two random samples of examinees. The differences between equatings of two random samples were slightly larger for the SPE pairwise equating than for the SPE EM method and the IRT method, especially at the extremes of the score range. This suggested that the SPE pairwise method may be less stable than the other two methods. However, these differences were too small to be of practical significance.

Studies of these two types of pre-equating models in general yielded satisfactory results when the tests involved a fairly large number of unscored (non-operational) items and used large random samples. However, only a few nationally administered admission and achievement tests have large enough populations of examinees and item

pools available for these types of analyses. Many professional licensing and certification tests are much shorter and have small examinee populations. The accuracy of equating methods depends on factors such as sample size, specific equating methods used, design used for data collection, and the nature of the tests to be equated. In actual testing practice, scores sometimes have to be made comparable under less-than-optimal conditions. Thus, it is helpful to know how well a specific equating method can perform under different conditions. The purpose of this study is to examine the adequacy of pre-equating models when certain sample and test characteristics are systematically varied.

The specific purpose of this study was to address the following questions which might be raised by a test publisher considering the use of an item chracteristic curve model or a linear section pre-equating model for achieving comparable scores.

1. When a pre-equating approach is applied to the equating of different forms, is there a practical difference in the accuracy due to the choice of methods?
2. How does the length of variable sections affect the accuracy of different pre-equating methods? And how does sample size of examinees affect the accuracy of pre-equating methods?

Design and Procedure

This study investigated the efficiency of pre-equating methods on simulation data. The simulation of item responses for examinees was accomplished using the modified three-parameter IRT model with a fixed

guessing parameter. The computer program DATAGEN (Hambleton & Rovinelli, 1973) was used to generate the item response scores for examinees. A large set of 144 items was generated and then split to form two half tests. Since these two half tests were created from items generated from the same distribution of item parameters, they are randomly parallel. One shorter test contained items 1 to 72, and the other shorter test contained items 73 to 144. These two parallel halves were arbitrarily assigned as operational form X and nonoperational form Y. Each of the two forms was then divided into sections of various lengths for study of the effect of test length.

The data analyses for the present study consisted of two basic phases: experimental and cross validation. The experimental phase involved equating the the scores on two tests by administering the first test to all examinees and parts of the second test to different samples of the same examinees. In this phase, a data set of 144 items and 3000 examinees was generated by computer, and samples of different size and test length were drawn from this large pool. A criterion equating based on the complete data set was established to evaluate individual equatings. The cross validation phase involved application of the derived procedure in the experimental phase to a cross validation sample of 3000 from the same population. The results were compared to the original criterion equating.

In both phases there were two methods (section pre-equating or SPE and IRT), two designs (one-variable-section test and two-variable-section test), three sample sizes (1500, 2250, and 3000) and five lengths (84, 90, 96 and 108 items for one-variable-section tests; 96,

108 and 120 items for two-variable-section tests). This design resulted in the 42 combinations were shown in Table 1 . Thus, each phase consisted of analyses on these 42 equating conditions.

Two random samples of 3000 observations from the same population were generated by separate computer runs in the experimental phase. The samples with smaller sizes (i.e., 2250 and 1500) were drawn from these sample. Pairwise present method was used for parameter estimations in 2VST SPE. For the 1VST SPE, correlations between the corresponding sections in Form X were borrowed for the estimation of covariances between variable sections. The IRT analyses were carried out in a concurrent design by the LOGIST program (Wingersky, Barton and Lord, 1982; Wingersky, 1983). The true-score equating method (Lord, 1980) was then applied to the parameter estimates to provide score equivalents for the two forms.

TABLE 1

Number of Sections, Test Length and Sample Sizes

(a) 2VST

Length of operational test	No. of section	Length of section	No. of samples	Total test length	Size of sub-sample	Total sample
72	3	24	3	120	500	1500
					750	2250
					1000	3000
72	4	18	6	108	250	1500
					375	2250
					500	3000
72	6	12	15	96	100	1500
					150	2250
					200	3000

(b) 1VST

72	2	36	2	108	750	1500
					1125	2250
					1500	3000
72	3	24	3	96	500	1500
					750	2250
					1000	3000
72	4	18	4	90	375	1500
					563	2252
					750	3000
72	6	12	6	84	250	1500
					375	2250
					500	3000

The Criterion Equating

Two kinds of comparisons were made: to see how well pre-equatings based on partial data worked as compared to complete data equatings, and also to see how individual pre-equating methods performed as they compared with each other. As we learn from previous studies, the equating results were most satisfactory when the assumptions of the particular method used were met. Since the data of this study were generated from a three-parameter logistic model with a fixed guessing parameter, the IRT three-parameter fixed-c method was chosen to be the criterion equating. In this criterion equating, raw scores of the original sample of 3000 on the 72 non-operational items (test Y) were equated to the raw scores of the same examinees on the 72 operational items (test X), and the method was supposed to produce the most "correct" equating among all the methods. If two pre-equating methods were used with the same groups and one of the methods produced the "correct" equating, then the transformed y score distribution on the scale of X would be expected to be closer to the transformed y score distribution on the criterion equating than the transformed y score distribution for the "incorrect" method to the criterion equating.

The accuracy of equating was assessed by the closeness of the transformed y score distribution to the criterion equating distribution for a particular method when both transformations were applied to the cross validation sample: (1) The criterion transformation, using modified three-parameter model and the total sample and total number of items was applied to the other sample of 3000; (2) The sample transformation for each method, sample size and number of items

was applied to the other sample; (3) A measure of error or difference between the two sets of scores for that equating was computed. The measure of error was defined as the standardized root mean square difference (SRMSD) of these two score distributions. This index takes the following form:

$$SRMSD_k = [\sum_{i=1}^{n} (x'_i - x'_{ci})^2 / s_c^2 n]^{.5},$$

where $SRMSD_k$ was the SRMSD for method K, x'_i was the transformed score for a particular score y_i by method K, x'_{ci} was its corresponding criterion equating score, s_c was the standard deviation of the criterion scores, and n equaled to 3000. Neither set of scores was reduced to integer values before the index was computed. The summation was over the cross validation sample which was also the replication sample. A double cross-validation design was used, and the two large samples of 3000 served as the cross validation sample for each other. Each equating was compared to the same standard, that is, the modified three-parameter IRT equating based on the complete data of its own sample.

Description of the Data Set

The items in the two test forms used in the present study were generated from the same set of 144 items. Form X, which was arbitrarily assigned as the "operational form", consisted of item 1 to item 72; and Form Y, which was assigned as the "pre-equating form", consisted of item 73 to item 144. Item difficulty parameters were selected at random from a uniform distribution in the interval [-2.0, 2.0]. The items had an estimated mean difficulty of 0.05 (SD=1.16,

ranging from -1.97 to 1.98). Item discriminations were selected at random in the interval [.5, 2.0] from a uniform distribution. The estimated mean discrimination was 1.24 (SD=0.40, ranging from 0.52 to 1.98). The constant guessing value was 0.20.

The two random samples of 3000 examinees were generated from the same population, and each individual examinee had a complete set of item responses for the total 144 items. When the items were split to form two tests, the average scores of the two forms were very close to each other. For Sample 1, the average score on Form X was 42.263, with a standard deviation of 12.759 and score range from 10 to 71; the average score on Form Y was 42.479, with a standard deviation of 12.467 and score range from 11 to 71. For Sample 2, the average score on Form X was 42.701, with a standard deviation of 12.799 and score range from 8 to 71; the average score on Form Y was 42.903, with a standard deviation of 12.453 and score range from 9 to 71.

Comparison of Equating Results

Results of IRT Pre-equatings

Table 2 shows the amount of SRMSD's for each method and the average SRMSD for each sample size across methods. The values listed in the table are the average values across all designs in two samples. For example, 0.0088 was the average over three-, four-, and six-section 2VST-IRT tests in two replications with sample size 3000. Table 3 gives the average SRMSD's for different test lengths and methods. The values presented are the overall average of two replications across all samples.

The same IRT analyses were repeated in two independent samples for each test length/sample size condition. Columns 3, 4 and 5 in Table 2 list the standardized root mean square differences for sample sizes of 3000, 2250 and 1500, respectively, when compared to the criterion equating. Two SRMSD values are presented for each individual combination: the upper figure was computed from Sample 1 and the lower figure from Sampel 2. The two criterion equatings were computed from the complete data sets of the two samples. The SRMSD's were calculated through the entire score range from 1 to 71.

TABLE 2

Average SRMSD for Each Method and Sample Size

(a) 2VST

Total test lengths	Sample	IRT Sample Size 3000	IRT Sample Size 2250	IRT Sample Size 1500	SPE Sample Size 3000	SPE Sample Size 2250	SPE Sample Size 1500
120	1	0	.0118	.0063	.0188	.0190	.0191
	2	.0042	.0069	.0104	.0195	.0194	.0197
108	1	.0086	.0112	.0080	.0205	.0213	.0191
	2	.0142	.0136	.0247	.0230	.0224	.0233
96	1	.0099	.0110	.0112	.0199	.0213	.0189
	2	.0159	.0204	.0240	.0225	.0222	.0257
Mean		.0088	.0125	.0141	.0207	.0209	.0210
SD		.0060	.0045	.0081	.0017	.0014	.0029

(b) 1VST

Total test lengths	Sample	IRT Sample Size 3000	IRT Sample Size 2250	IRT Sample Size 1500	SPE Sample Size 3000	SPE Sample Size 2250	SPE Sample Size 1500
108	1	.0174	.0175	.0376	.0212	.0242	.0270
	2	.0089	.0107	.0209	.0271	.0204	.0269
96	1	.0169	.0192	.0221	.0267	.0259	.0277
	2	.0208	.0176	.0080	.0282	.0253	.0192
90	1	.0277	.0215	.0347	.0362	.0299	.0391
	2	.0192	.0303	.0400	.0209	.0214	.0317
84	1	.0262	.0314	.0378	.0334	.0310	.0387
	2	.0366	.0361	.0463	.0271	.0227	.0236
Mean		.0217	.0230	.0309	.0276	.0251	.0292
SD		.0084	.0086	.0127	.0053	.0038	.0070
Over-all Mean		.0162	.0185	.0237	.0246	.0233	.0257
SD		.0098	.0088	.0137	.0054	.0036	.0069

Despite the individual differences between the two sample replications for the individual conditions, consistent effects of sample size and test length can be observed. In all except three cases, for any test length, the amount of SRMSD increased when sample sizes decreased from 3000 to 2250, and from 2250 to 1500. When the SRMSD's were averaged over the lengths of test, the effect of sample size became more obvious.

The effect of test length was also observable in the SRMSD's. In general, for all sample sizes, SRMSD's increased as the test became shorter. Table 3 shows that, when averaged over sample sizes, the mean SRMSD's increased as the test length decreased in all except one case. In Table 2 we can find that the average SRMSD's for two-variable-section tests (2VST) are much larger than the average SRMSD's for one-variable-section tests (1VST). This difference can be attributed to the differences in test lengths because the one-variable-section tests generally had fewer items than the two-variable-section tests. The overall means listed at the bottom of Table 2 were the average of the mean SRMSD's of 1VST and 2VST equatings. For those two-variable-section and one-vaiable-section tests containing equal numbers of items, the average SRMSD's of 1VST were found to be larger than the average SRMSD's of 2VST. These results are also listed in Table 3

TABLE 3

Average SRMSD over Sample Size

Length	IRT 2VST	SPE 2VST	IRT 1VST	SPE 1VST	IRT Average	SPE Average
120	.0048	.0192	-----	------	-----	-----
108	.0117	.0216	.0188	.0245	.0153	.0230
96	.0154	.0217	.0174	.0255	.0164	.0236
90	-----	-----	.0289	.0298	-----	------
84	-----	-----	.0357	.0294	.0357	.0294

Results of SPE Methods

The last three columns in Table 2 present the standardized root mean square differences between each SPE analysis and the criterion equating. The changes in the amount of SRMSDs associated with changes in sample size were less significant for the SPE than for those of the IRT pre-equatings. Only two cases showed increasing SRMSD's when sample sizes reduced. In some cases the changes in SRMSD's were irregular, while in other cases the SRMSD's almost remained the same regardless of the changes in sample size. The average SRMSD's of the 2VST equatings were similar for sample size of 3000, 2250 and 1500, but the SRMSD's for sample size 1500 had larger variability. For 1VST equatings, sample size 2250 had the lowest SRMSD's in half of the cases. Discrepancies for sample size 2250 also had lowest overall mean and standard deviation.

The effect of test length on SPE SRMSD's can be found in both Tables 2 and 3 . Most SRMSD's increased as the number of items in the

tests decreased. We can see in Table 3 that the average SRMSD's for three different sample sizes also increased with the decrease in lengths of the tests, and the only inconsistency was found at the length of 84 items. For tests of 108 items and 96 items, the one-variable-section tests had higher average SRMSD's than the two-variable-section tests over all three sample sizes.

Comparison of SPE and IRT Pre-equatings

The important results of the comparison between the SPE and IRT pre-equatings are summarized as follows:

1. Across equating designs and lengths, the average SRMSD of the SPE method was greater than the average SRMSD of the IRT method across all sample sizes (ie, 3000, 2250, and 1500).
2. Across test designs and lengths, the average SRMSD increased as the sample size decreased for the IRT methods in both 2VST and 1VST pre-equatings, while the average SRMSD values were relatively stable across sample sizes for the SPE method in 1VST pre-equatings.
3. At each level of sample size, the average SRMSD for the SPE method was greater than the average SRMSD for the IRT method, and this difference decreased as the sample size decreased.
4. In general, across all designs and sample sizes, for both the IRT and SPE methods, the longer the test, the smaller the average SRMSD. For tests of equal lengths (e.g, 108 and 96 items), the 1VST pre-equatings had larger average SRMSD's than their 2VST counterparts.

5. The average SRMSD for the SPE method was greater than the average SRMSD for the IRT method when total test lengths were 120, 108, 96, and 90 items. This difference became smaller when the number of test items decreased. As the number of items dropped to 84, the average SRMSD for the SPE method became smaller than the average SRMSD for the IRT method.

All of the trends listed above were based on the average of two-sample replications. Irregularities were found in individual cases. Comparison of equating results within each method across three sample sizes (i.e., 3000, 2250, and 1500) confirms the stability and consistency of the equatings. Take the 1VST-2 section test as an example: the X score equivalents were plotted against the integer Y raw scores for the IRT and SPE methods in two data sets. Figures 1 through 4 show the relationships between converted X scores obtained from equatings using different sample sizes, converted scores from the criterion equating, and the Y raw scores. The conversion lines obtained from the three samples were very close to the conversion line from the criterion equating, especially at Y scores range between 30 and 55, where two-thirds of the scores were centered.

Three smaller sample sizes (900, 600 and 300) were also used for the 84-item test. The SRMSD's are listed in Table 4 . The IRT equatings for sample sizes of 600 and 300 were not very reliable because the LOGIST runs failed to converge. However, we can still see the effect of sample size on those equatings. As sample size reduced, the amount of SRMSD increased, the IRT equatings became much worse than the SPE equatings for those very small samples. The conversion

lines obtained from these small sample SPE and IRT equatings are presented in Figures 5 through 8. Comparison between the two methods showed that the small size IRT equatings diverged more from the criterion in the middle range of scores than did equatings by the SPE methods. The SPE conversion lines, when compared to the IRT criterion conversion line, tended to underestimate scores at the lower extreme and overestimate scores at the higher extreme, but were very close to the criterion in the middle of the score range.

TABLE 4

SRMSDs for Very Small Samples in A 84-item Test

Sample size	IRT Sample 1	IRT Sample 2	SPE Sample 1	SPE Sample 2	Mean
900	.0305	.0597	.0400	.0192	.0374
600	-----	-----	.0374	.0225	.0410
300	-----	-----	.0350	.0411	.0696
Mean	-----	-----	.0375	.0276	

Discussion of Results

This study attempted to examine equating results obtained from applying the SPE and IRT pre-equating methods to the same set of data. The equating based on complete data has provided a criterion for pre-equatings using partial data. Closeness of the pre-equating results to the results of criterion equating served to evaluate the equating models and designs. Results from this study show that, under

the conditions of computer generated data, there is little difference between methods. This was especially true when both test length and sample size were large. Although the differences were small, they tended to favor IRT, which is expected since the data were generated from an IRT model. These differences, however, are probably of more theoretical than practical significance because most of the standardized root mean square differences in equating are less than 0.04. For practical equating situations where scores are rounded to the nearest whole numbers, those differences even become less significant. In fact, the largest differences found between the criterion and individual equatings are only two points in terms of raw X scores. They were in the samples of size 300 for both SPE and IRT equatings.

The use of simulated data is probably the reason why the differences between pre-equating results and the criterion were so small. The data sets were generated from a modified three-parameter logistic model; therefore, they fitted perfectly the IRT analysis. When compared to a criterion based on the same model, the differences between results from the IRT pre-equating and IRT "post-equating" criterion can be treated totally as differences between equating using partial data and equating using complete data. On the other hand, the SPE method is designed to fit the data with a linear conversion. The differences between the SPE results and the criterion also reflected the differences between linear conversion and IRT conversion. These differences occurred at the extremes of the score scale, where smaller numbers of cases were found. In real test situations, problems such as position effects of variable sections (e.g., practice or fatigue),

unequal difficulty of the same items to pre-equating samples and operational samples, and violation to the model would result in less satisfactory equatings by the IRT method. Larger differences between criterion and pre-equatings and across individual pre-equatings would be expected if real data had been used.

The primary focus of this study was the effects of sample size and test length on equating results. The effects of these two factors were not exactly the same for the two equating methods. The effect of sample size was greater on the IRT pre-equatings than on the SPE methods. For the IRT equatings, the SRMSD's increased with the decrease in sample size. For the SPEs, SRMSD's were essentially stable with the changes in sample size. The effects of test length were found in both SPE and IRT methods. The average SRMSD's across all sample sizes increased when the number of items in the tests decreased. Again, the IRT equatings tended to be more sensitive to the change in test length than the SPE methods. The unsatisfactory results for IRT pre-equatings in small samples can be attributed to the inaccuracy in parameter estimations because the LOGIST program failed to converge for sample sizes 600 and 300. This problem resulted in constant overestimations in the middle range of scores where most scores were located. Although the SPE results tended to underestimate scores at the lower end and overestimate scores at the upper end--and such a tendency was even stronger when sample size became smaller--they provided better conversions for the middle part of the score range than did the IRT results.

The comparisons between 2VST and 1VST designs show that for a given test length, 2VST usually had smaller SRMSD's than 1VST in the same condition (i.e., same method and sample size). It was expected that for SPE, 2VST would produce better equatings than 1VST because 2VST estimated the covariance of variable sections from sample data while 1VST had to borrow the correlations from operational sections for the covariance estimation. But in this study, Form X was so similar to Form Y that the correlations between sections in Form X were excellent approximations for the correlations in Form Y. Since the 2VST equatings were also better than the 1VST equatings for IRT methods, the borrowed correlations seemed not to be the only cause that accounted for this phenomenon. One possible explanation is that the difference between 2VST and 1VST was attributed to the difference in variable sections and number of subsamples. The variable sections for 2VST and 1VST had different items for subsamples. Using the test of 108 items as an example, for a total sample size of 1500, the 2VST had six subsamples of size 250 while the 1VST had two subsamples of size 750. It could be that many small samples were able to yield more information than a few large samples in equating.

The results of this study confirmed those of Leary and Wightman (1985) that, for relatively large sample sizes and long tests, the IRT method provided better equatings than the SPE method. The most important finding of this study is that the differences between SPE and IRT pre-equatings became closer as the test became shorter and sample size became smaller. The superiority of IRT to SPE disappeared gradually as a result of decrease in sample size and test length.

Notice that when the test had only 84 items, the SPE results were better than the IRT results. When sample size dropped below 1000 for the test of 84 items, this reverse tendency continued. The SRMSD's for IRT equatings became extremely large at the sample size of 300. Discrepancies also increased for SPE results in very small samples, but at a slower rate.

Conclusion

Based on the analyses in this study, the results tend to support the following general conclusions:

1. Given an equating sample size over 1500 and data that fit a modified three-parameter logistic model, both the IRT and the SPE methods provided excellent equating results. However, the IRT method has greater flexibility in test construction and administration because its focus is at the item level rather than test level. IRT allows items from various sources to be pre-calibrated separately and pooled together, while SPE would have all the pretest items given in a single administration. IRT also allows test takers of different levels of competence to have different parts of the test while SPE requires equivalent samples for stable parameter estimation.
2. The IRT method seemed to provide more consistent parameter estimations than the corresponding SPE method, but it required a considerably large sample size and number of items. Some trade off may be reasonable between sampling requirements and accuracy at the lowest and highest ends of the score scale.

3. The parameter estimation in the IRT method became unstable for smaller sized samples, and the results of the IRT pre-equating were less satisfactory than those of the SPE method.
4. Of the two equating methods included in this study, the SPE method is simpler and less expensive than the IRT method and produced results judged to be at least as stable and consistent as the IRT method when sample size and test length were small (in this study, when sample size was under 1000 and items were fewer than 90). For this reason, the SPE method may be the method of choice when small sample sizes are used for horizontal equating of shorter tests.

In summary, for a pre-equating situation in which the samples are small and the tests are relatively short, the SPE method can be recommended as a viable procedure. For longer tests with larger samples of examinees, the IRT method is a procedure of better precision.

Recommendations for Future Research

As mentioned earlier, this study has several limitations. First, only simulated test data were used. These data satisfied all the requirements for both SPE and IRT models. For real test data, we shall always expect some violation to the models. Some human factors that may occur in real test situations such as practice effects and fatigue were also not considered in this study. All the comparisions were conducted under perfect conditions. Studies using real test data should be conducted to verify the results of this study.

Also, similar studies might be carried out with a modified one-parameter logistic model, since the estimation procedure used for this model works better with smaller sample sizes than that used for the modified three-parameter model. Perhaps this approach has some advantage over the SPE method in small samples.

Except in the case of tests of 84 items, the sample sizes did not go below 1500. However, the results of this study seem to suggest that SPE started to gain superiority over IRT when the sample size went below 1500. In order to draw a clearer conclusion, more comparisons involving smaller samples and fewer test items should be conducted.

One question left unanswered is why in this study the results of 2VST equatings were always better than those of 1VST equatings of the same test length. The 2VST and 1VST tests are not directly comparable in this study because they had different items for variable sections. A direct comparison between those two designs should use the same items for variable sections, making the only difference between 2VST and 1VST the number of variable sections that the examinees are given.

REFERENCES

Angoff, W. H. (1984). Scales, Norms, and Equivalent scores. Princeton, N. J.: Educational Testing Services.

Cook, L. L. & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate test. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 175-195). Vancouver, B.C.: Educational Research Institute of British Columbia.

Cowell, W. R. (1982). Item-response theory pre-equating in the TOEFL testing program. In P. W. Holland and D. B. Rubin (Eds.), Test equating. New York: Academic Press, Inc., 149-162.

Hambleton, R. K., & Rovinelli, R. (1973). A FORTRAN IV program for generating examinee response data from logostic models. Behavioral Science, 18, 74.

Holland, P. W. & Thayer, D. T. (1984). Section pre-equating in the presence of practice effects. Princeton, N.J.: Educational Testing Service, Technical Report No. 84-44.

Holland, P. W. & Wightman, L. E. (1982). Section pre-equating: A preliminary investigation. In P. W. Holland and D. B. Rubin (Eds.), Test equating. New York: Academic Press, Inc., 271-297.

Leary, L. F. & Wightman, L. E. (1985). A compaison of SPE and IRT equating results obtained from verbal and quantitative aptitude test data. Paper presented at the Annual Meeting of the American Ecudational Research Association (Chicago), April 1985.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, N. J.: Educational Testing Services.

Wingersky, M. S. (1983). LOGIST: program for computing maximum likelihood procedure for logistic test models. In Hambleton, R. K. (Ed.) Applications of item response theory (pp. 45-56). Vancouver, B.C.: Educational Research Institute of British Columbia.

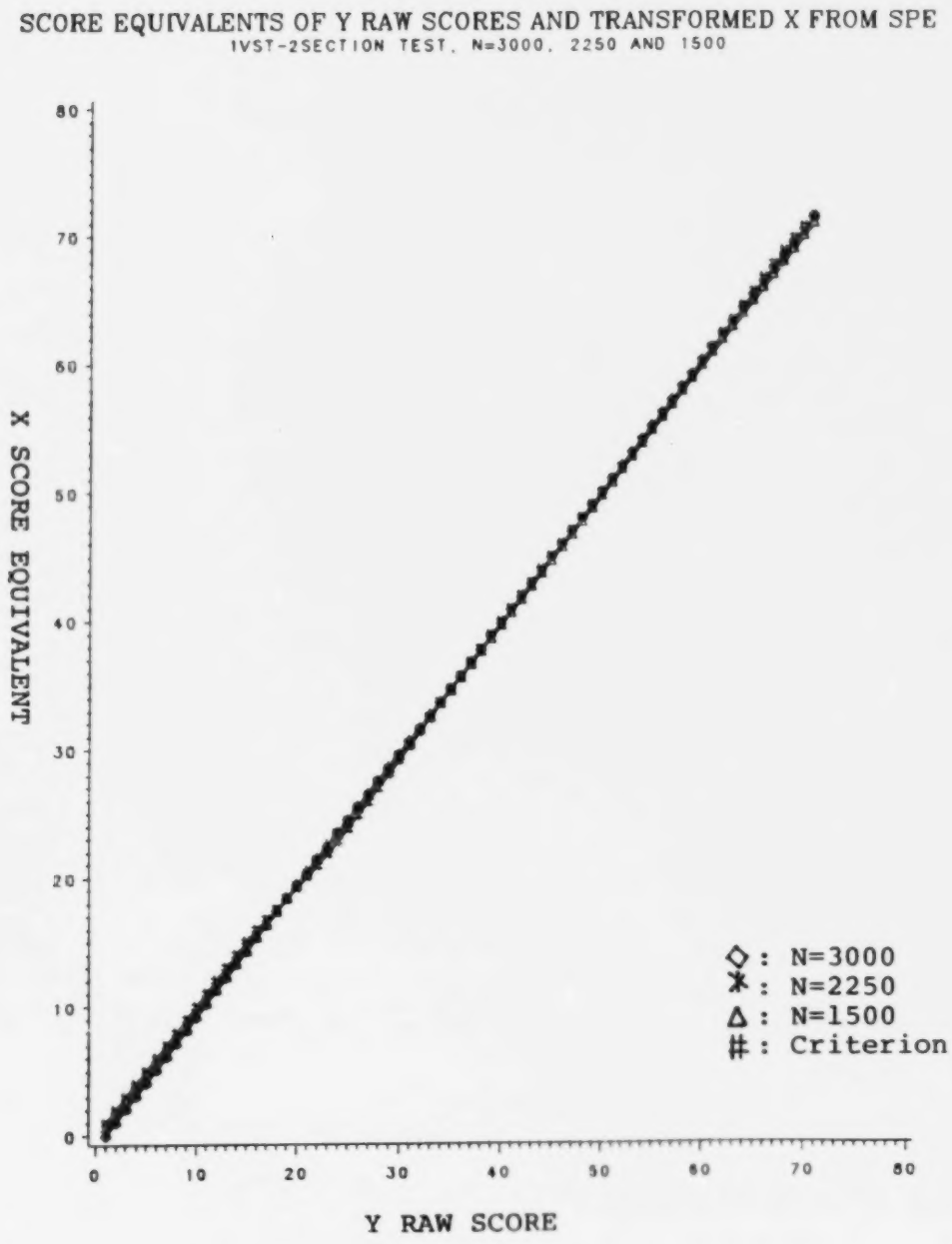

Figure 1. Sample 1 Score Equivalents for a 108-Item Test Obtained from SPE

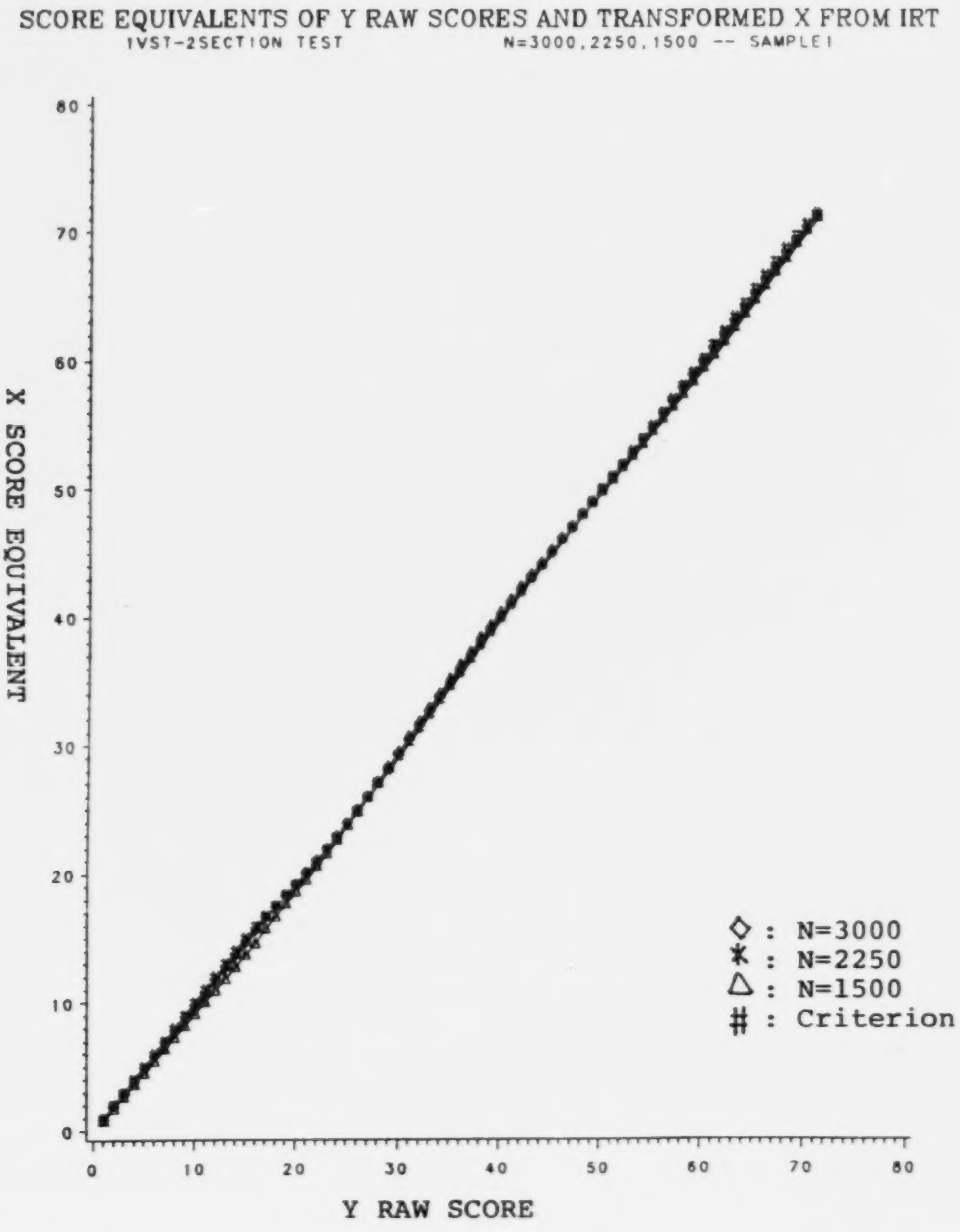

Figure 2. Sample 1 Score Equivalents for a 108-Item Test Obtained from IRT

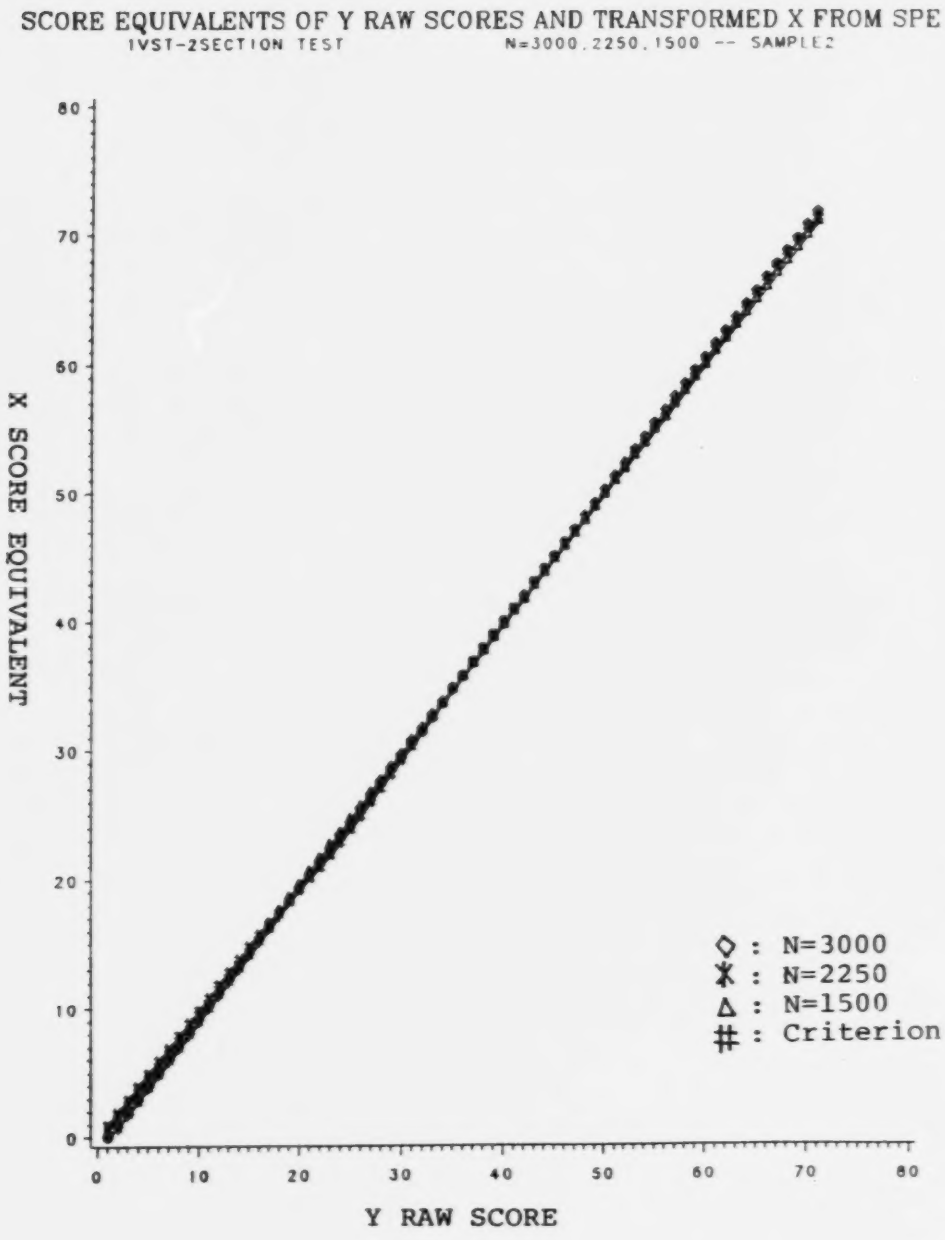

Figure 3. Sample 2 Score Equivalents for a 108-Item Test Obtained from SPE

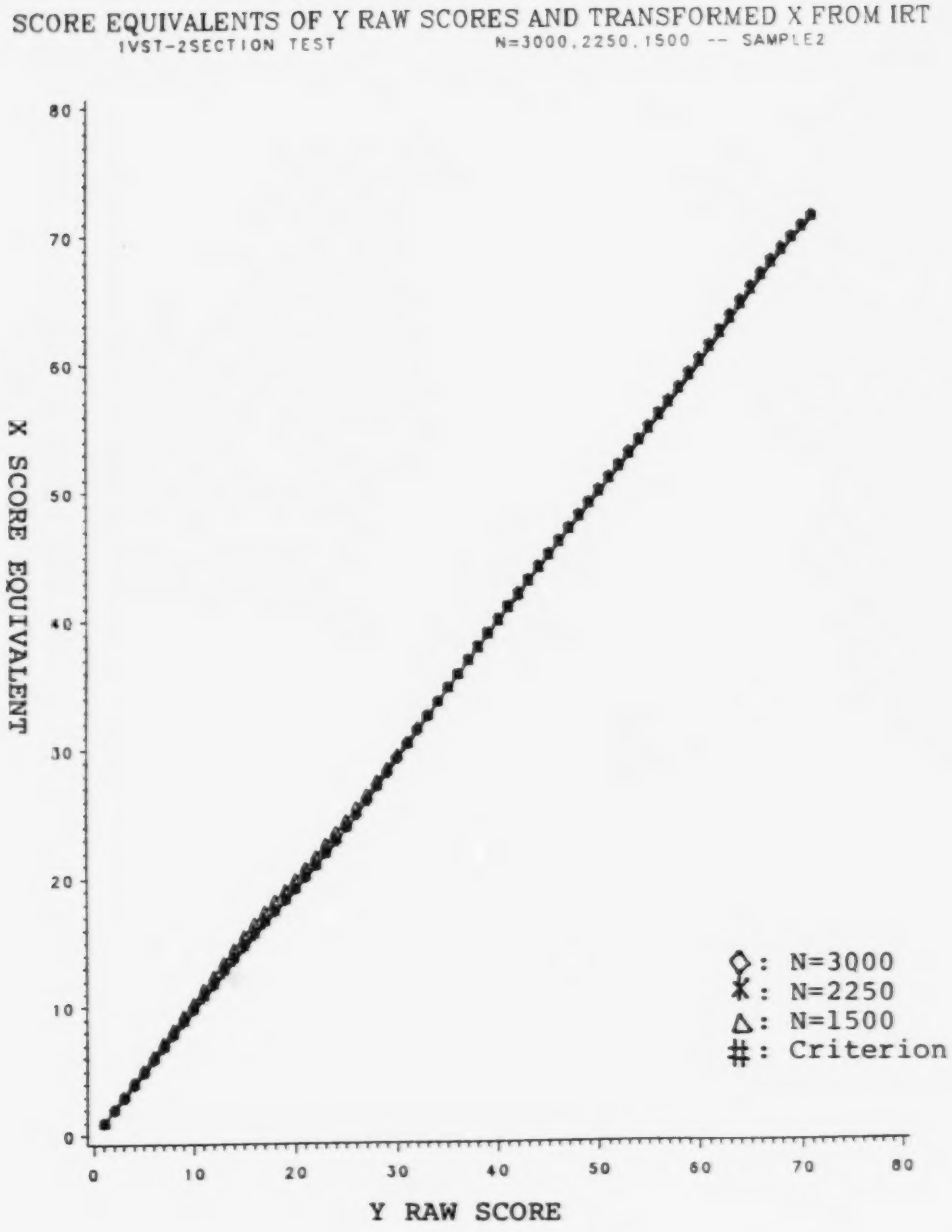

Figure 4. Sample 2 Score Equivalents for a 108-Item Test Obtained from IRT

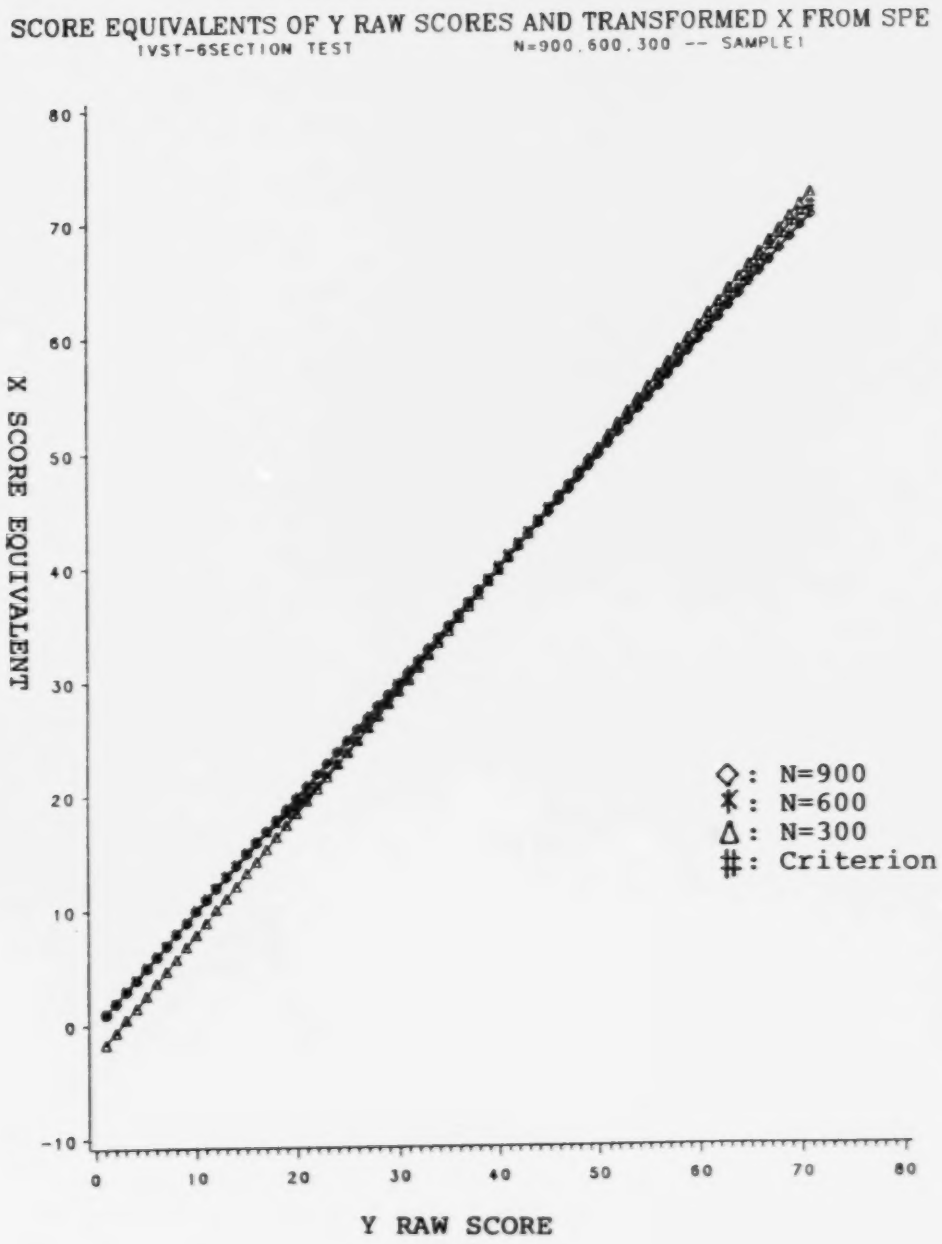

Figure 5. Sample 1 Score Equivalents for a 84-Item Test Obtained from SPE

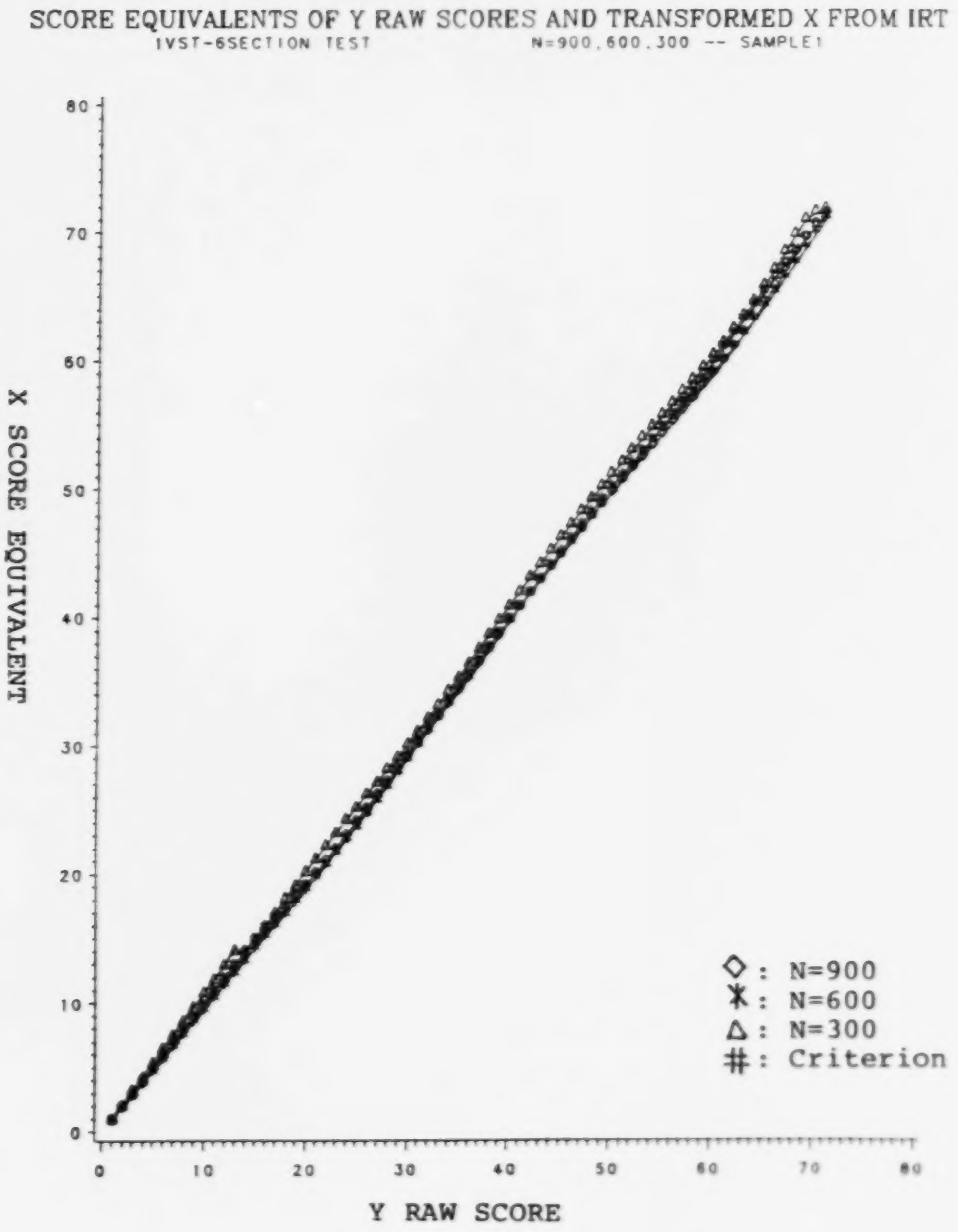

Figure 6. Sample 1 Score Equivalents for a 84-Item Test Obtained from IRT

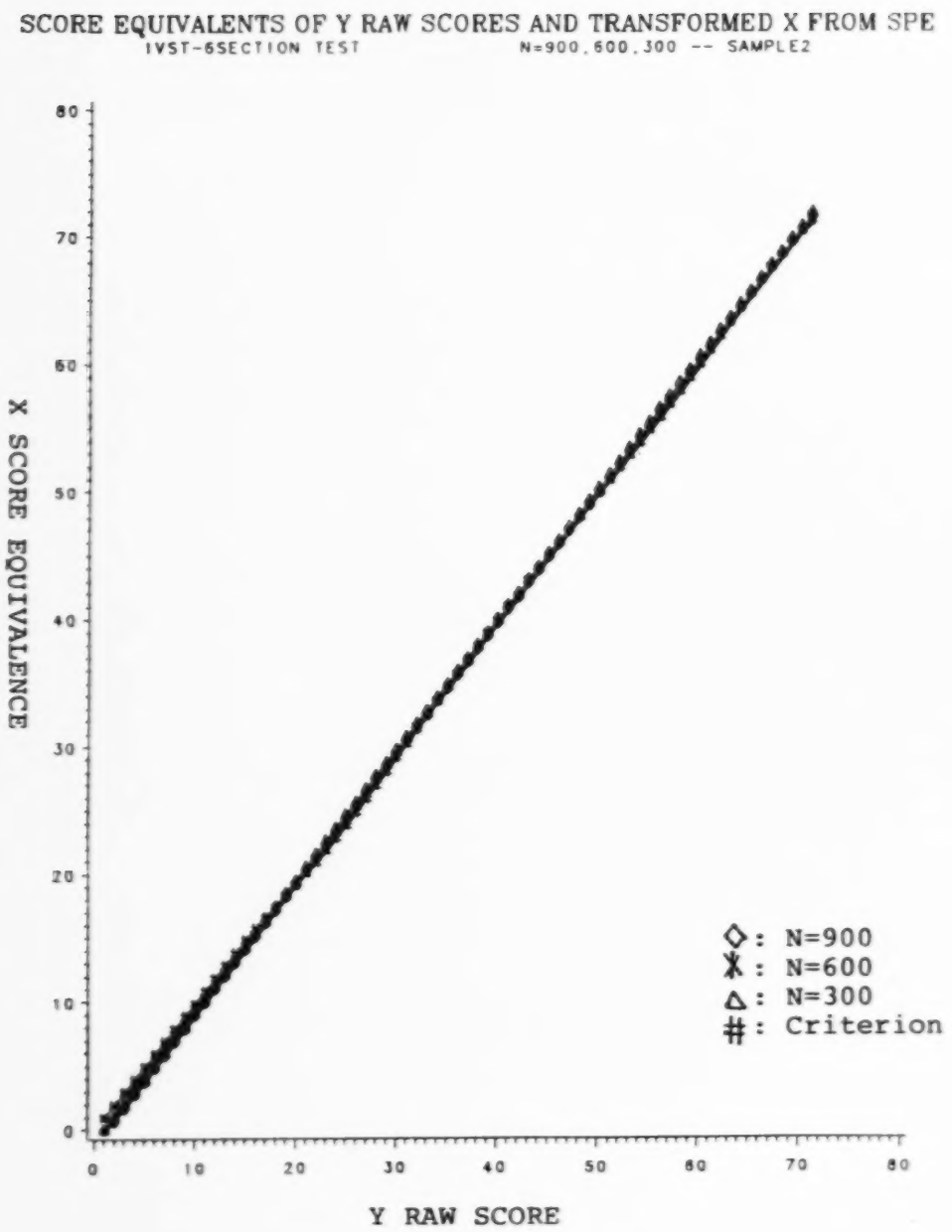

Figure 7. Sample 2 Score Equivalents for a 84-Item Test Obtained from SPE

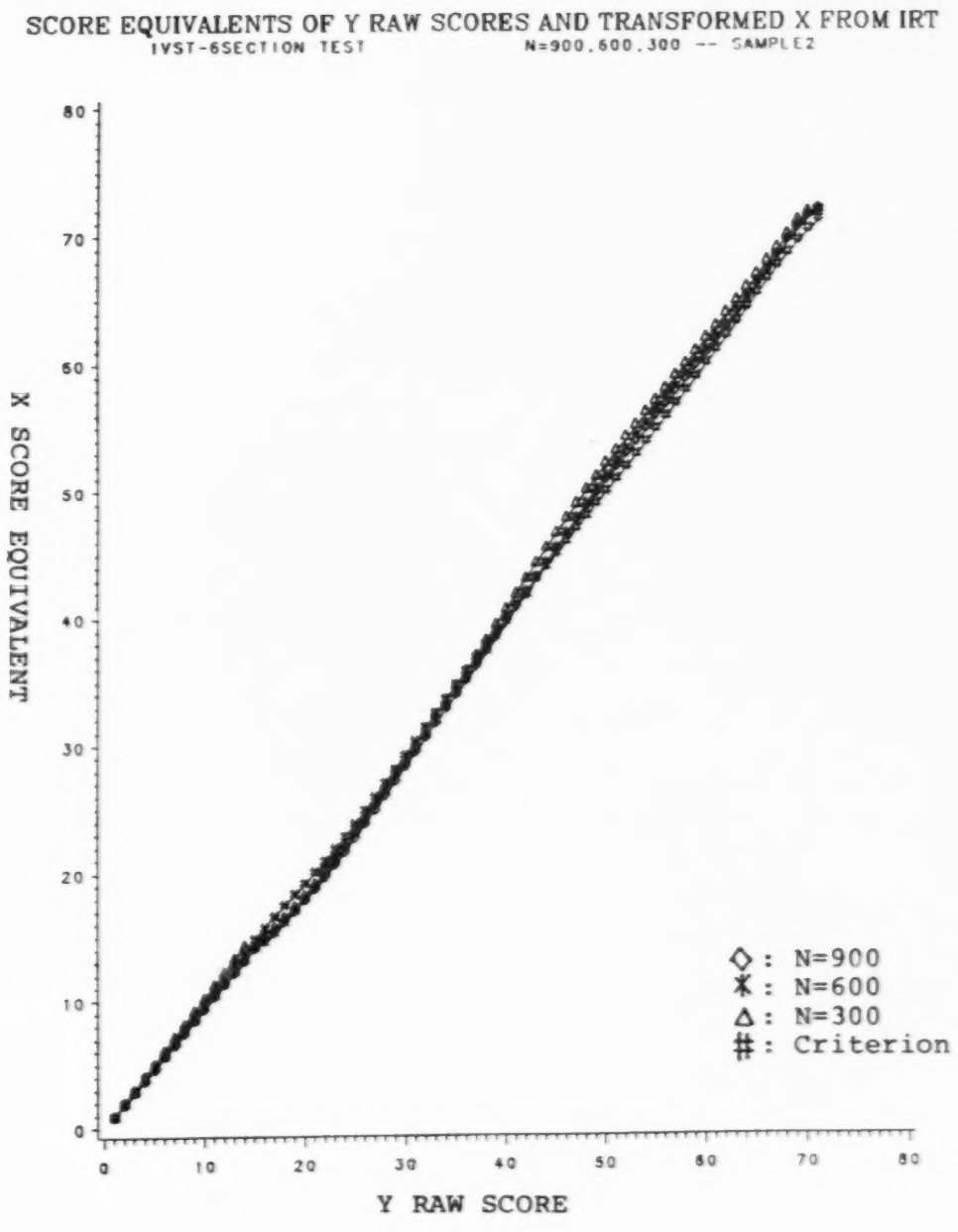

Figure 8. Sample 2 Score Equivalents for a 84-Item Test Obtained from IRT